Remarks

The Examiner is thanked for the courtesy shown during the telephone interview on February 22, 2005. During the interview the Applicants' representative explained that the specification erroneously identified SEQ ID NO: 2 as the coding sequence for the *dep* protein. Similarly, SEQ ID No. 1 is erroneously described as containing the coding region. In fact, SEQ ID NOs: 1 and 2 include the coding region for *ydhC*. As shown in Fig. 3 and discussed, for example, on page 9, lines 13-23 and page 11, lines 21-25, *ydhC* is distinct from *dep*.

To clarify, the *dep* protein has the amino acid sequence of SEQ ID NO: 3. Neither SEQ ID NO: 2 or SEQ ID NO: 1 include the coding sequence for SEQ ID NO: 3. A BLASTP alignment between SEQ ID NO: 3 and the amino acid sequence encoded by SEQ ID NO: 2 is enclosed. The alignment report indicates that "[n]o significant similarity was found". This amendment corrects the inadvertent error in the specification.

The amendment introduces no new matter. The disclosure of the appropriate amino acid sequence (SEQ ID NO: 3) of a protein is adequate written description and renders definite claims directed to a gene that encodes the protein. *In re Wallach*, 71 U.S.P.Q. 2d 1939, (Fed Cir 2004).

It is the Applicants understanding that the previous rejections under 35 U.S.C. §§ 102 and 103 were based on the false premise that the protein of SEQ ID NO: 3 was encoded by the *ydhC* gene. Because the error in the specification and resultant misunderstanding have been corrected, it is requested that the rejections based on the *ydhC* gene be reconsidered and withdrawn.

The *ydhC* gene is identified in Gen Bank record Accession No. AE000261, which was purportedly submitted by Blattner in connection with the 1997 *Science* article cited in previous Official Actions. It is the Applicants understanding that the Examiner has taken the position that a

plasmid having the complete coding sequence for *ydhC* was inherently disclosed in the reference because Blattner allegedly necessarily created such a plasmid in the course of sequencing the complete genome of *E coli*.

It is hereby noted that SEQ ID NO: 3 is identical to the amino acid sequence described as the translation of gene *b1657* in Gen Bank record Accession No. AE000261. The product of *b1657* is merely described as a putative transport protein. It is respectfully submitted that Blattner (in combination with Accession No. AE000261) neither inherently discloses, nor renders obvious, a plasmid having a gene that codes for the *dep* protein.

Blattner does not inherently disclose a plasmid having a gene encoding *dep*

In order for a reference to inherently anticipate a claim element, the element must necessarily be present in the reference. MPEP § 2112. The fact that a certain condition may be present in the prior art is insufficient to establish inherency. *In re Oelrich*, 212 USPQ 323, 326 (CCPA 1981).

The activities of Blattner did not necessarily create a plasmid that includes a gene encoding the *dep* protein. Although Blattner identified *b1657* as a gene coding for the amino acid sequence of SEQ ID NO: 3, it is respectfully submitted that Blattner did not necessarily create a plasmid containing the complete sequence of *b1657*. According to the record for Accession No. AE000261, the *b1657* gene is located between *sodB* and *purR*, which are found between positions 1,700,000 and 1,800,000 on the *E. coli* genome (*see* the map foldout of Blattner). According to Blattner (page 1453, column 3, lines 24), this region of the genome was sequenced using the M13 Janus shotgun strategy. This method is described by Burland, *et al* (1993) Genome sequencing on both strands: the Janus strategy, *Nucleic Acids Res* 21(15):3385. The Janus strategy article was coauthored by Frederick Blattner and a copy is enclosed.

According to the Janus strategy, samples are sonicated for 50 seconds to randomly shear DNA into fragments of from a few hundred base pairs to several kilobase pairs. The fragments are size-fractionated, and fragments of 0.7-2 kb are collected and ligated into the vector for sequencing. The majority of the fragments collected are length 1-2 kb. *See*, page 3386, column 1, paragraphs b and c of Burland, *et al.*

According to the record for Accession No. AE000261, *b1657* is 1170 bp (about 1.2 kb) in length. In the process of randomly creating 1-2 kb fragments by sonification, there is a significant chance that no single fragment would contain any particular 1.2 kb segment. In fact, it is not even possible for a segment of this size to fit into fragments at the lower end of the range of fragments selected and cloned in the Janus shotgun method. Therefore, there is a significant chance that the Janus shotgun strategy used by Blattner produced no random fragment that contained the entire 1170 bp sequence of *b1657*. It follows that the Blattner sequencing method did not necessarily produce a plasmid having the entire sequence of *b1657*. Therefore, the Applicants respectfully submit that Blattner did not necessarily create a plasmid that would have the entire *b1657* sequence or otherwise encode for a protein that confers DHCP resistance. The mere possibility that such a gene might have been created is not adequate to support an inherency rejection. See *Oelrich* at 326.

Blattner does not render obvious a plasmid having a gene encoding *dep*

It would also not have been obvious to create a plasmid that includes a gene encoding for the DHCP resistance gene, *dep*, at the time the invention was made. A proper obviousness analysis requires the difficult but critical step of casting the mind back to the time the invention was made. *In re Dembiczak*, 50 USPQ2d 1614, 1616-17 (Fed Cir 1999). At that time, *b1657* was merely identified as a putative transport protein, one of 146 such proteins identified by Blattner. Blattner provides no

evidence that *b1657* is expressed in *E. coli*, no reliable basis to believe that it encodes a transporter if it is expressed, no description of what *b1657* might transport if it is a transporter, and certainly no suggestion that *b1657* would be useful for conveying resistance to DHCP.

At best, the disclosure of *b1657* as a putative transport protein could be characterized as general guidance to perform further research on the gene. This suggestion establishes a classic "obvious to try" situation. In such cases, a general disclosure might "pique a scientist's curiosity, such that further investigation might be done as a result of the disclosure, but the disclosure itself does not contain a sufficient teaching of how to obtain the desired result , or that the claimed result would be obtained if certain directions were pursued." *In re Eli Lilly & Co.*, 14 USPQ2d 1741, 1743 (Fed Cir 1990). Just as discussed in *Eli Lilly*, given the description of "putative transport protein" a scientist might be inclined to further research the role of *b1657*. It is a long-standing rule that this type of "obvious to try" scenario does not render the invention obvious. *In re Deuel*, 34 USPQ2d 1210, 12616 (Fed Cir 1995).

Blattner does not contain any teaching as to how to find a gene that confers DHCP resistance. In addition, there is no suggestion that a gene conferring DHCP resistance could be found if known techniques of molecular biology were employed. Finally, Blattner does not disclose or suggest that *b1657* is a DHCP resistance gene. In other words, based on the teachings of Blattner, one skilled in the art would have no reasonable expectation of success in creating a plasmid that encodes for a DHCP efflux protein based on the mere description that *b1657* is a putative transport protein.

A similar issue was decided by the Board of Patent Appeals and Interferences in *Ex Parte Obukowicz* 27 USPQ2d 1063 (Bd Pat App & Int 1992). In that case, the applicant was claiming a method of combating plant insect pests by applying a plant-colonizing bacteria having within its

chromosome heterologous DNA encoding for the protein toxin of *Bacillus thuringiensis*. The examiner issued a rejection based on a primary reference that described incorporation of the claimed gene into plasmids of various bacteria. The primary reference further suggested incorporation of the gene into the plant itself to create a systemic toxin or into specific bacteria that have superior survival characteristics.

In reversing the examiner's rejection under 35 USC § 103, the Board characterized these statements in the reference as "an invitation to scientists to explore a new technology that seems a promising field of experimentation" and "of the type that gives only general guidance". Further, the Board reiterated the long standing rule that these types of general suggestions may make an approach "obvious to try", but do not make the invention obvious. 27 USPQ2d at 1065.

Further, the Board observed that given the teachings of the applicants' specification regarding incorporation of the gene into the bacterial chromosome and using the bacteria in the plant environment, one can explain the rationale for the invention using selected teachings from the prior art. This approach, however, was recognized as impermissible hindsight. *Id.*

The present situation falls squarely into the same category as the "obvious to try" cases. At best, Blattner provides general guidance on genes in the E coli genome, and possibly even a motivation to scientists to perform further research on the 146 putative transport proteins disclosed therein. However, there is nothing in the reference that give one skilled in the art a reasonable expectation of success in creating a plasmid that contains a gene encoding a DHCP efflux protein. Therefore, the invention is non-obvious over Blattner.

Claim Rejections under 35 U.S.C. § 112 ¶ 2

Claims 16-18 have been rejected as allegedly indefinite based on the term "disrupted" as used

in connection with the *pur*, *ydhC* or *ydhB* genes. Claim 17, relating to a disrupted *ydhC* gene, has been cancelled. The term "disrupted" has been clarified in claims 16 and 18 by reciting the restriction enzymes that can be used to "disrupt" the *purR* and *ydhB* genes, respectively. One skilled in the art would readily be able to determine the relevant restriction sites (MluI and NruI or Eco47III, respectively) within the each gene. Given the new recitations, one skilled in the art would be able to determine the sites at which the genes at issue would be cleaved or "disrupted" without undue experimentation. As such, the recitations are clear and definite.

Claims 16-19 have been objected to based on the terms "*pur* gene", "*ydhC* gene" and "*ydhB* gene". Specifically, the Official Action indicates that the nature of these genes is unclear. The recitation of "*pur*" has been amended to "*purR*" in claims 16 and 19. With respect to all three terms, a claim term is definite if one skilled in the art would understand the scope of the claim when read in light of the specification. *Union Pacific Resources Co. v. Chesapeake Energy Corp.*, 57 USPQ2d 1293 (Fed Cir 2001). The specification clearly describes the genes in question at page 11, lines 21-25. *PurR* encodes purine synthesis repressor. *YdhB* encodes a homologue of the cyn operon transcriptional activator. The y*dhC* gene encodes a homologue of bicylcomymycin. As such, the applicants respectfully submit that one skilled in the art would understand the nature of these genes in light of the specification.

Claim Rejections under 35 U.S.C. § 112 ¶ 1 – Written Description

Claims 16-19 have been rejected as allegedly failing to meet the written description requirement. In this respect, the Official Action indicates that the specification provides an adequate written description for the genus of plasmids encoding *dep*, but fails to envision certain subgenera defined by the absence of specific genes or the inclusion of disrupted forms of those genes.

The objection to claim 16 specifically relates to disrupted forms of the *purR* gene. As noted above, this claim has been amended to further define the subgenus to include the *purR* gene disrupted by the restriction enzyme MluI. It is respectfully submitted that the specification envisions the full scope of this subgenus.

The objection to claim 18 relates to disrupted forms of the *ydhB* gene. This claim has been amended to recite that the *ydhB* gene disrupted by NruI or Eco47III. The specification envisions the full scope of this subgenus as well.

Claim 19 defines a subgenus based on its independence from *purR*, *ydhC* and *ydhB*. Claim 19 has been subordinated to depend from new independent claim 21 (as have claims 16 and 18). As noted in the Official Action, the application includes support for the subject matter of the new independent claim (the genus of plasmids having a gene that encodes *dep*). As a dependent claim, claim 19 properly further limits claim 21. Moreover, the narrower scope of claim 19 is fully supported by the specification, such as in the embodiment of pSP007, into which *dep* was cloned without *purR*, *ydhC* and *ydhB*. Because the cloning of genes into plasmids is routine in the art, it is respectfully submitted that no further examples are necessary to describe the subgenus of dependent claim 19.

Claim Rejections under 35 U.S.C. § 112 ¶ 1 – Enablement

Claim 20 has been rejected as allegedly lacking an enabling disclosure. Specifically, it is indicated that one skilled in the art must use pSP007 to practice the invention. As indicated in the Applicants' response of May 11, 2004, pSP007 was deposited with the American Type Culture Collection, 10801 University Boulevard, Manassas, Virginia 20110-2209, on September 5, 2001 and assigned Accession No. PTA-3682. The specification has been amended to reference the deposit in

compliance with 37 C.F.R. § 1.809(d). Further, it is hereby certified that all restrictions on accessibility to pSP007, on deposit with the American Type Culture Collection under Accession No. PTA-3682, will be irrevocably removed by the Applicants upon the granting of a patent.

New claims

New claims 21-25 are directed to plasmids and isolated nucleic acid molecules comprising the *dep* gene. These claims are patentable for the reasons set forth above, *i.e.*, they are neither inherently anticipated nor rendered obvious by Blattner. Support for the new claims can be found throughout the specification, for example, at page 11, line 26 through page 12, line 8.

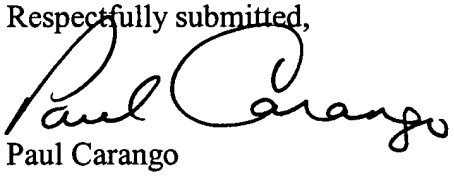
New method claims 26 and 27 have been added to the application. These claims are directed to methods of using plasmids of the invention in anticipation of rejoinder once claims directed to the plasmids are allowed. Support for the method claims can be found, for example, in original claim 14 and on page 8 of the specification

Conclusion

The application has been amended to correctly reflect that SEQ ID NOs: 1 and 2 include the coding sequence for *ydhC* and not for *dep*. Therefore, it is respectfully requested that the rejections under 35 U.S.C. § 103 based on Blattner be reconsidered and withdrawn. It is further requested that similar rejections not be established based on *b1657* because Blattner neither inherently discloses nor renders obvious a plasmid having a gene that encodes for a DHCP efflux protein. It is also requested that the rejections and objections under 35 U.S.C. § 112 be withdrawn for the reason set forth above.

It is respectfully submitted that the entire application is now in condition for allowance, which action is respectfully requested. If the Examiner believes that minor amendments or attention to other matters of form will advance the case, the Examiner is invited to telephone the Applicants' undersigned representative.

Respectfully submitted,



Paul Carango
Reg. No. 42,386
Attorney for Applicants

PC:SAN:vbm
(215) 656-3381

Blast 2 Sequences results

PubMed Entrez BLAST OMIM Taxonomy Structure

BLAST 2 SEQUENCES RESULTS VERSION BLASTP 2.2.10 [Oct-19-2004]

Matrix BLOSUM62 gap open: 11 gap extension: 1
x_dropoff: 50 expect: 10.000 wordsize: 3 Filter Align

Sequence 1 lcl|seq_1 **Length** 389

Sequence 2 lcl|seq_2 **Length** 403

No significant similarity was found

```
CPU time:      0.02 user secs.       0.01 sys. secs          0.03 total secs.

Lambda      K       H
   0.326     0.138     0.399

Gapped
Lambda      K       H
   0.267    0.0410     0.140


Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Sequences: 1
Number of Hits to DB: 1111
Number of extensions: 740
Number of successful extensions: 1
Number of sequences better than 10.0: 0
Number of HSP's better than 10.0 without gapping: 0
Number of HSP's gapped: 1
Number of HSP's successfully gapped: 0
Number of extra gapped extensions for HSPs above 10.0: 1
Length of query: 389
Length of database: 791,805,785
Length adjustment: 132
Effective length of query: 257
Effective length of database: 791,805,653
Effective search space: 203494052821
Effective search space used: 203494052821
Neighboring words threshold: 9
Window for multiple hits: 0
X1: 15 ( 7.0 bits)
X2: 129 (49.7 bits)
X3: 129 (49.7 bits)
S1: 40 (21.7 bits)
S2: 77 (34.3 bits)
```

Nucleic Acids Research, 1993, Vol. 21, No. 15 3385–3390

Genome sequencing on both strands: the Janus strategy

Valerie Burland, Donna L.Daniels, Guy Plunkett, III and Frederick R.Blattner
Laboratory of Genetics, University of Wisconsin, 445 Henry Mall, Madison, WI 53706, USA

Received May 8, 1993; Revised and Accepted June 15, 1993

ABSTRACT

The design of large scale DNA sequencing projects such as genome analysis demands a new approach to sequencing strategy, since neither a purely random nor a purely directed method is satisfactory. We have developed a strategy that combines these two methods in a way that preserves the advantages of both while avoiding their particular limitations. Computer simulations showed that a specific balance of random and directed sequencing was required for the most efficient strategy, termed the Janus strategy, which has been used in the *Escherichia coli* genome sequencing project. This approach depended on obtaining sequence easily from either strand of a cloned insert, and was facilitated by inversion of the insert in the engineered M13 vector Janus, by site-specific recombination. The inversion was accomplished simply by growth on the appropriate host strain, when the DNA strand incorporated into the new single stranded phage was complementary to that in the original phage, and was sequenced by the same simple protocol as the first strand.

INTRODUCTION

Large scale DNA sequencing projects such as whole genome analysis demand special attention to strategy: which is more efficient, to sequence random clones from a library or to sequence specific clones from an ordered set in a directed manner? The scale of genome projects also demands consideration of technical simplicity (suitability for automation), informatics requirements and cost, all vital in designing the most efficient system. Construction of random libraries allows sequencing to proceed without prior finely detailed mapping or specific subcloning steps, but a high order of redundancy is necessary for a randomly sequenced project to be assembled with no specific 'finishing' steps. Different directed strategies, on the other hand, yield sequences which fit economically together in a predetermined fashion, but require a significant level of analysis or design in the form of mapping work to identify a set of overlapping template subclones, or in designing new primers for sequencing and DNA amplification in the case of primer 'walking'. Clearly both these strategies have valuable features. In the *Escherichia coli* genome project (1, 2) we have combined the particular benefits of each, using random sequencing to collect the bulk of the data initially, followed by a simplified directed strategy that uses no new clones or technical procedures, to complete the sequence. Computer simulations were used to investigate the effects of different combinations of random and directed data collection on cost and efficiency, and the results indicated the optimimum scheme. Seamless integration of the two strategies was facilitated by a bacteriophage M13 vector, Janus, engineered to allow sequencing of both strands from single stranded templates. A single cloning process yields Janus library clones whose inserts may be inverted simply by propagation on a host expressing a site specific recombinase, Int from phage lambda (3). The packaged genome then incorporates the second strand of the insert, and the same simple protocol may be used for template preparation. Sequencing random Janus clones followed by selection of specific clones for sequencing after inversion leads to a particularly cost-effective strategy striking a balance between random and directed approaches.

MATERIALS AND METHODS

Recombinant DNA techniques

Standard methods (4) were used for the construction of Janus with few modifications. Enzymes were used according to the manufacturers instructions and were obtained from Boehringer-Mannheim (restriction enzymes), United States Biochemical (Sequenase, Mung bean nuclease), New England Biolabs (T4 DNA ligase), and Stratagene (alkaline phosphatase). The *attP* fragment was obtained from pPH54 (5) provided by A.Landy. The fragment was isolated by gel electrophoresis through 2% Low Melting Point agarose (BRL) in Tris-acetate buffer, and was eluted using Geneclean (Bio101 Inc.). Bacterial strains were obtained from New England Biolabs (JM101) and Stratagene (XL1-blue).

Oligonucleotides

The *attB* oligonucleotide strands (sequences are shown in Fig. 3A) and special sequencing primers were synthesised by the Beckman Research Institute of the City of Hope. A primer was designed for ease of sequencing across the *attB* part of the Janus candidates to confirm the structure of both the original and inverted forms, and also to analyse false positives in the libraries. This primer hybridised within the *lacZ* gene about 100 bases 3′ of the cloning site; the primer sequence is 5′ CCTCTTCGC-TATTACGCC 3′.

Construction of random libraries

a) *Growth and preparation of DNA for cloning*. For large DNA preparations, Janus was grown in NZY medium on host JM101; phage were collected from 2 liter culture supernatants by precipitation with 4% PEG, 0.5MNaCl, and banded on cesium

chloride gradients. After dialysis to remove CsCl, DNA was released by phenol extraction and then dialysed extensively. The double stranded replicative form of Janus was prepared as a plasmid by alkaline lysis (4) and purified by cesium chloride density gradient centrifugation. Lambda clones were grown in 2 liters of NZC, using host strain LE392, and MOI= 0.01 (6). Phage were collected by PEG precipitation and purified by cesium chloride banding; DNA was prepared from the phage as for single strand Janus.

b) Sonication and size fractionation of target DNA. 5μg of lambda clone DNA were adjusted to 20μl with TE buffer in 0.5ml microfuge tubes. The samples were sonicated for 50 seconds using a cup-horn probe, with fresh ice cold water at the same level for each sample. Power was delivered by a Branson Sonifier (Cell Disruptor W-350) set at 17% of duty cycle and output control at 1, giving an output of 200W. These settings normally sheared all of the phage DNA, giving a size range from 2−300 hundred basepairs (bp) to about 10 kilobasepairs (kb), evenly distributed. The sonicated DNA was then repaired by digestion with 15 units of Mung bean nuclease, at room temperature for 10 minutes. The sample was then immediately loaded onto a 1% Seaplaque GTG agarose gel (FMC BioProducts) in Tris-acetate buffer and electrophoresed for 25 minutes at 25 volts. The parts of the gel containing the marker lanes were cut off and stained with ethidium bromide. The gel was then reassembled and the 0.7−2 kb sonicated DNA cut out and eluted by Geneclean (Bio101 Inc.), into 20 μl TE buffer; the yield was 0.5−1.0 μg. Samples in which sonication produced an uneven distribution of material in the 0.7−2.0 kb range were not used as they often contain an excess of very small fragments which clone more efficiently than the 1−2 kb pieces desired.

c) *Cloning with Janus.* In the *Escherichia coli* genome project (1, 2) Janus was used to make subclone libraries starting from *E.coli* clones in phage lambda (7). The following procedure was adapted from Bankier *et al.* (8). The vector was prepared by digestion with Sma I and treated with alkaline phosphatase to prevent recircularization. Target DNA was prepared from lambda clones by sonication to produce fragments ranging in size from 0.5 to several kb. Target DNA ends were repaired by digestion of single strand extensions with Mung Bean nuclease; other enzymes such as Klenow polymerase or T4 polymerase, or combinations of these, all give similar results. After end repair, the target DNA was size-fractionated by agarose gel electrophoresis, and fragments from 0.7 to 2kb were collected. Purified fragments were ligated into the prepared vector and transformed into XL1-blue competent cells prepared by the method of Hanahan (protocol 2 in reference 9). When these transformed bacteria were plated, JM101 was used to provide the bacterial lawn. This combination of host strains permitted the maximum efficiency of transformation and good growth of plaques on agar, both of which are vital for optimal recovery of library clones and fast throughput in a large project. Clones containing lambda DNA inserts were identified and discarded before sequencing.

To test the quality of prepared vector batches and monitor the cloning process, a preparation of lambda DNA was partially digested with Alu I. After fractionation to prepare the same size range as the sonicated targets, it was ligated into the vector sample. Since Alu I leaves blunt ends, no repair step is necessary and the cloning efficiency is at least 10 fold higher than that of the end-repaired target.

The yield of clones (colorless plaques) was between 1000 and 3000 per 100ng prepared target DNA. In comparison with enzymatically digested DNA, mechanically sheared DNA clones inefficiently because many of the ends are unrepairable and fragments may also be internally damaged. However, all the restriction enzymes we tested singly or in combination (10) did not produce truly random libraries and mechanical methods are preferred. In the Janus libraries the ratio of clones to vector (colorless to blue plaques) was about 2 or 3 to 1, and the background of false positives (colorless plaques from a vector-only ligation) was 2 or 3 per 100 blue. These blue plaques derive from uncut or unphosphatased molecules and the few false positives result from loss of bases at the Sma I site.

Screening for lambda in the Janus libraries

Library plaques were toothpicked into 100μl TE buffer in microtiter dish wells. Each dish was then replicated onto a lawn of bacteria (JM101) in top agar and incubated 18 hours at 37°, producing the microtiter dish array of phage samples as clear patches of lysis. The set of patches were lifted onto Nytran membranes (Schleicher and Schuell), denatured and neutralized, then baked at 68° for at least 2 hours. The membranes were hybridized to a lambda probe made by labelling Hind III-cut DNA by nick translation with ^{32}P. Standard procedures were followed (4). The membranes were exposed to Kodak XAR5 film and the hybridizing patches readily identified; these are marked in the microtiter dish wells with food color and the remaining samples transfered to a new dish. After they have been used to inoculate mini-cultures for sequencing, the phage samples are solidified by addition of 75 μl of melted 0.4% agarose in TE buffer, cooled, sealed and stored at 4°.

DNA sequencing

For the large scale sequencing project, Janus library clones were grown and DNA template preparation carried out in microtiter dishes (11) for random sequencing. To analyse candidates for the Janus construct, templates prepared from 1ml cultures were extracted and purified in microfuge tubes (12). DNA sequencing was by the chain termination method of Sanger (13) and was performed by an automated system, using internal ^{35}S label, acrylamide gel electrophoresis and autoradiography (14). Reagents and Sequenase were obtained from United States Biochemicals.

Inversion of Janus inserts

The inverting host strain was constructed by transforming JM101 with the plasmid pHS3-1 provided by J. Gardner (15), containing the cloned lambda *int* gene fused to the P_{tac} promoter. The transformed strain, called FB1898, was maintained on minimal agar with ampicillin (100μg/ml). For inversion of selected Janus clones on this host, bacteria were grown in NZY broth containing ampicillin, to mid-exponential phase, then 2ml portions were inoculated with the Janus samples. IPTG was added to 0.1mM and the cultures incubated at 37° with shaking for 6 hours to give titers similar to those obtained on JM101 without the plasmid. Phage was harvested and DNA templates prepared as for noninverted clones. In tests of Janus plaque color, spontaneous inversions and reversions were not detected; the frequencies of both were less than 1 in 10^5.

Software

Computer programs for data collection, sequence assembly, and sequence analysis (1) were from DNASTAR packages. The program for the strategy simulations was specially written.

RESULTS

Strategy of sequencing with Janus

For each lambda clone, sequences were collected from the random Janus library until sufficient data were obtained to allow assembly (alignment) of sequences into contigs of several kb in length. For a library made from a 20 kb lambda clone, this number is about 500 sequences depending on the length of sequence readouts and data quality. These data assembled into 3–4 contigs having an average coverage of 6–8×, with some areas of only 1 or 2× coverage and other areas with data from only one strand. To address all these problem areas, specific Janus clones were selected by inspection of the alignment. Taking into account the average insert length, clones were chosen so that after inversion, new sequences collected from the opposite strand and the opposite end of the inverted clones would cover the problem areas. The new data extended the ends of contigs enabling them to merge, as well as improving coverage to meet the minimum standard of at least four determinations at every point and at least one on each strand, essential for the level of accuracy needed to locate authentic open reading frames and other features. This strategy has been used in the *E.coli* genome project and is illustrated in Fig. 1, a detailed example from one of the sequenced lambda clones (2) where sequencing inverted clones improved a poorly covered area.

Simulations of combined strategies

To assess the optimum mix of random and inverted sequencing, a theoretical analysis was carried out. The effect of switching at different times from random data collection to selection of specific clones for inversion was analysed using a computer program to compare different finishing methods for simulated assemblies of random sequences. The simulation calculated the cost of finishing a 20 kb project by varied amounts of random sequencing followed by primer walking or Janus inversions to finish. Collection and assembly of raw data were determined by random number generation with assumed values for data quality, sequence readout length and cost of each operation. The minimum criteria for coverage require at least two determinations on each strand at every point. The number of finishing steps required (gaps or thin areas) was captured at different times during the simulated assembly. The curves in Fig. 2 show the cost of a sequencing project using different strategies: cost was plotted as a function of the amount of random sequence collected before finishing by a directed method. The upper curve represents finishing solely with primer walking steps, and the lower curve shows finishing with as many Janus inversions (flips) as possible followed by the much smaller number of primer walking steps needed. In this case, after collecting random sequences to each fold of coverage, the computer examined the assembly for places where Janus clones may be flipped and sequenced from the second strand to advantage. Finally a few primers were used to close the sequence. In either case the result was coverage at least twice on each strand at every point.

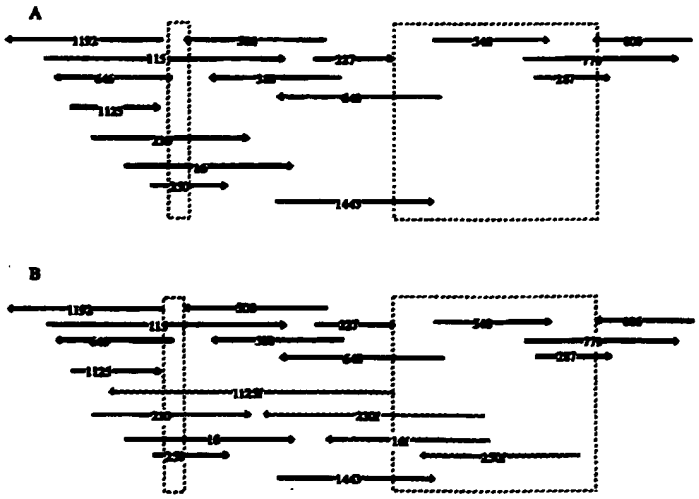


Figure 1. Sequencing strategy diagram generated by Seqman alignment computer program, edited for clarity. Data from the *E.coli* genome project (2); sequenced lambda clone Ec 27-236 (at 83.5 minutes on the physical map), a small portion of the lambda project is shown. Numbers represent Janus clone names. A: alignment of sequences collected in the random phase of the project, boxed areas are poorly covered or have data from one strand only. Clones 16, 230, 250 and 1125 were selected to provide additional data by inversion, and sequence was gathered from the opposite end for the complementary strand as described. B: alignment with the new sequences added (hatched arrows), showing improved coverage and second strand data for most of the problem areas. The two sequences from before and after inversion may overlap depending on the clone length and the length of the sequence. Data overlapping the flanks of the region shown were removed for clarity.

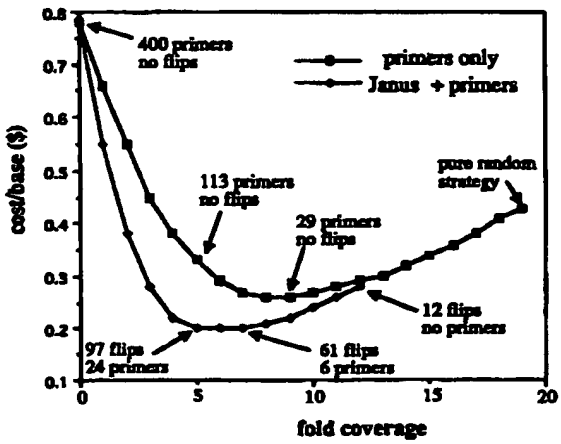


Figure 2. Comparison of costs of sequencing strategies with and without Janus inversions (flips) by computer simulation. Data points are averages of 25 simulated assemblies. Strategies using different combinations of primer walking and random sequence collection with no inversions (open squares), upper curve; with inversions, (filled diamonds), lower curve. Cost per finished base ($) was plotted against average number of determinations at each point by random data collection (fold coverage). Cost per finished base included the fixed costs of labor and materials in the current project for template preparations, sequencing reactions, gels and autoradiography but did not include management, research or other overhead. Assumed values are: cost per raw base, 2.45 cents; total cost per gel (24 sequences) at 85% yield and 400 bases per readout, $200; cost of primer for each walking step, $30; primer spacing 200 bases; average length of insert in Janus clones, 1 kb.

The zero fold point (no random sequencing) is a pure primer walking strategy. At 5 fold coverage without flips, the number of primers is reduced more than 90% and the cost to finish more than halved. At 19 fold coverage, no primers are required and the strategy is pure random. The minimum cost for finishing by primer walking is at 8–9 fold random coverage, when 30–40 primers are needed. Using the Janus strategy, the minimal cost is between 5 and 7 fold random coverage. At 7 fold, 61 very cheap inversion steps reduce the need for new primers from 58 (without flips) to 6. These data are summarized in Table 1. Thus the point to switch to flipping is at 5 to 7 fold coverage for optimum efficiency, achieved with a 25% saving in cost. Above 10 fold the costs are equivalent and increase linearly as the burden of obtaining very large quantities of data outweighs the smaller number of finishing steps needed.

We also tested a number of other assumptions. A fourfold reduction in the cost of primers reduced the cost of the pure primer walking strategy by half, whereas a tenfold drop in primer cost produced a saving of two thirds. At 100 fold reduction in primer costs, such as might be obtained by the use of pentamer, hexamer or nonamer libraries (16, 17, 18), the cost per base for pure primer walking was reduced four fold. This does not take into account the high initial cost or the data management challenge. As primer costs become negligible there is little difference between the two mixed strategies by this simulation. At this point other considerations become more important.

Design and construction of Janus

M13 cloning vectors (19, 20) have the proven advantage of producing single stranded templates for primer extension (13) without competition from the complementary strand. Since M13 grows without lysing the host bacteria, DNA may be isolated easily by harvesting bacteriophage from culture supernatants, free from cell proteins, RNA or host genomic DNA. In addition the vector is distinguishable from phage containing cloned inserts by screening for plaque color (20). Janus was constructed from M13mp19 so that these features were preserved. Obtaining sequence from the second strand of the cloned insert would normally entail recloning in the opposite orientation or extensive random sequencing of shotgun clones to obtain both strands by chance. The second strand could be obtained by enzymatic synthesis on the first strand template (21), or by purification of the double stranded replicative form from cell lysates, procedures which lose the advantage of the simple preparation protocol for single strand DNA. By engineering a recombination system to invert the insert during growth, the second strand is made available for packaging and preparation as single strand template.

The site-specific recombination system for lambda integration into the *E.coli* genome was chosen for its efficiency and because it is well understood. The lambda and *E.coli att* sites by which the phage integrates into the genome (*attP* and *attB* respectively) have been analyzed in detail (reviewed in 3) and the functional elements defined. Both *att* sites consists of a 15 basepair common core where the staggered cut sites for the Int recombinase are located. The *attB* site has a few bases on either side of the core which are essential for activity (22). The *attP* site is much larger, with essential flanking sequences of 160 and 82 basepairs, containing binding sites for the Int recombinase and IHF protein (5). Lambda integration is a reversible event but a second lambda-coded gene product (Xis) is required for excision. This gene is not present in the Janus construct, thus only one recombination event can occur. The hybrid *att* sites resulting from inversion are not recognised by Int so the inversion product is stable. The design of Janus is shown in Fig.3A.

To construct Janus, the two *att* sites were engineered into M13mp19 with a unique cloning site between them. Their relative orientations were designed to invert rather than delete the cloned segment in an intramolecular recombination. Inversion is a site-specific recombination event and acts upon the double stranded DNA of the replicative form of the phage. The whole region

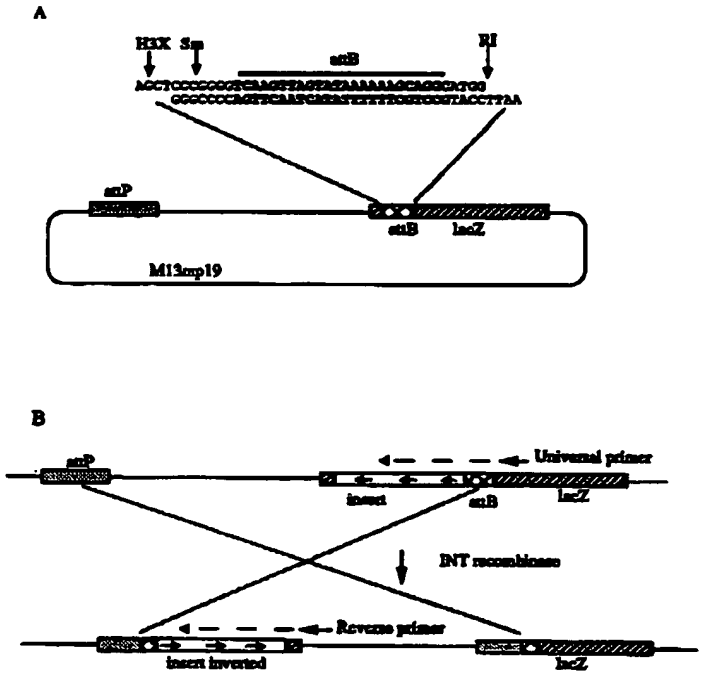


Figure 3. Structure of Janus and inversion of the cloned insert. A) Diagram of the vector showing the relative positions of *attB* and *attP* segments and the single cloning site Sma I. Sequence of the synthetic segment is shown, with the *attB* core in boldface type; H3X indicates the 5' extension designed to ligate to a cut Hind III end but not regenerate the Hind III site. At the other end, RI indicates the Eco RI-complementary extension. The restriction site was regenerated but is not useful for cloning since it is on the wrong side of *attB* to obtain sequence by inversion. The dashed arrow represents sequence obtained from the Universal primer. B) Inversion of the cloned insert in the presence of Int recombinase. Before inversion sequence is obtained by extension of the Universal primer (dashed line above the upper section). The arrows in the insert box indicate the orientation of the insert with respect to the primer site. Upon inversion, recombination junctions are formed by the fused halves of *attB* and *attP* at each site. Sequence is now obtained by extension of the Reverse primer, indicated by the dashed arrow.

Table 1. Summary of cost comparisons for different strategies

Strategy	Cost[a]	Random coverage	Number of Primers	Total coverage
Primer walking only	0.78	0	400	4
Random only	0.43	19	0	19
Mixed[b]	0.26	8–9	29–42	9.2
Janus[c]	0.19	6–7	6–13	8.1

[a] $ per finished base
[b] Random plus primer walking
[c] Random plus Janus inversions plus primers

between the *att* sites, including the cloned insert and the Reverse primer site, is inverted relative to the origin of replication (Fig.3B) so that the strand now packaged into phage heads is that complementary to the packaged strand of the original clone. The Reverse primer site is also on the packaged strand of the inverted DNA and is appropriately placed for sequencing the inverted insert (Fig.3B). Reverse sequencing primers are commercially available. Molecules not inverted (about 3 percent measured by plaque color of the inverted vector) do not interfere with the sequencing reactions since these do not contain the Reverse priming site.

Screening for plaque color in M13 depends on expression of the β-galactosidase α-complementing segment coded by part of the *lacZ* gene. Insertion of foreign DNA into the cloning site located in the exon disrupts the gene and the enzyme is no longer produced (19). An *attB* segment was designed to read in frame after insertion into the *lacZ* gene. The segment was synthesized with an adjacent Sma I site for library cloning (Fig.3A). Single strand extensions of Hind III and Eco RI restriction sites were used to ligate the segment between these sites in M13mp19. The *attP* site was obtained as a 400 basepair Hind III-Bam HI fragment from plasmid pPH54, a subclone of the *att* region from lambda (5). The ends of the purified fragment were filled in using Klenow polymerase and ligated into the Ava II site (similarly filled in) in M13mp19. Thus the two *att* sites were introduced into M13mp19 with the unique Sma I site between them. The construct's ability to produce blue plaques on plates containing X-gal was confirmed. Since the *attP* fragment was ligated by a blunt end reaction, it was important to determine the orientation of *attP* within the construct. Each *att* insertion was identified by hybridization of candidate plaques to ^{32}P- end labelled or nick translated *att* probes. Candidates which formed blue plaques and hybridized with both probes were tested for the ability to invert by a functional assay. Colorless and blue plaques were counted after growth in the presence of induced Int recombinase; inversion dissociates the *lacZ* promoter from the coding sequence and prevents expression of β-galactosidase. Correctly oriented constructs gave 100% blue plaques before inversion, and yielded more than 95% colorless plaques when plated 2 hours after Int induction in liquid culture. The structure of Janus was then confirmed by restriction site mapping and by sequencing across the *attB* segment in the uninverted and across the 3′ recombination junction in the inverted form (Fig.3B).

DISCUSSION

Tackling the analysis of entire genomes has called for new ways of combining technically simple random sequencing with some level of directed approach for the finishing steps. In the *E.coli* project, the key to integrating random and directed sequencing strategies was the ability to invert the insert within the vector. Obtaining sequence from the second strand of selected clones without having to make double stranded template was a significant advantage and enabled the mixed strategy to be developed. Now project finishing is possible with a relatively small number of easy directed steps and no further mapping or subcloning, while the order of redundancy needed is much reduced.

Purely directed strategies have the advantage that a much smaller amount of data is collected than in random projects, and that the fit of each new sequence to the existing data is tailored. The success of most directed strategies however, depends on precise identification and analysis of the template clones before sequencing: for example, nested deletions (23) and sequencing from mapped transposons (24, 25) entail prior characterization of the clones. Although there are strategies that avoid the work of subcloning and mapping, such as 'multiplex' sequencing (26) or primer 'walking', the design of oligo primers or probes from sequence at the end of a runout is a potentially limiting factor. Since new oligonucleotides must be designed and synthesized for each sequencing reaction, good design depends on having good data at the extreme end of the readout and long readouts are necessary to keep the number of primers to a minimum—both of course possible but difficult to sustain in a high throughput project. Failed reactions must always be repeated since all the sequences in the strategy plan are necessary for completion. In addition many starting points must be used simultaneously for closure in a reasonable time, creating significant costs of sample tracking and record keeping. In walking strategies, multiple hexamer priming (16, 17, 18) may reduce primer costs significantly. Such methods may eventually be able to use genomic DNA directly as a template, though this has yet to be demonstrated even by 'cycle' sequencing with a thermostable polymerase (reusing the template in thermal cycles) (27). Thus either some subcloning (28) is required, or the target DNA must be amplified from the genome by the Polymerase Chain Reaction (29) before sequencing. Finally, if these methods are used to produce sequence with only one determination at each point they become vulnerable to mutations in the template and to sequencing errors, not all of which are easily detected.

The advantages of random sequencing are that no analysis of the template clones is necessary, that only one process is needed to collect the data, and that success does not depend on the success of any individual sequence reaction, but depends only the total amount of data collected. Informatics requirements are therefore modest. These factors become particularly important when a large-scale project is undertaken. However, purely random strategies demand prohibitively large amounts of data for completion and accuracy. For example, to obtain coverage of each base at least twice on each strand, average coverage must approach 20×, or 1760 clones for a 20 kb segment compared to fewer than 600 by the Janus stategy or 400 by primer walking only. Janus takes advantage of the conceptual simplicity of the purely random strategy but reduces the amount of effort involved; at the same time the finishing steps are acquired at nearly the same cost as random data.

ACKNOWLEDGEMENTS

This is paper 3348 from the Laboratory of Genetics. Development, construction and verification of Janus was supported by the Romnes Fund through the University of Wisconsin Graduate School. The *E.coli* Genome Project is supported by NIH award HG00301. We thank M.Schwid, B.Fritz and E.Sommers for technical assistance, J.Gardner and A.Landy for kindly providing plasmids and N.Peterson for administration. We also thank the members of the *E.coli* Genome project teams.

REFERENCES

1. Daniels,D.L., Plunkett III,G., Burland, V. and Blattner,F.R. (1992) *Science*, 257, 771–778.
2. Burland,V., Plunkett III,G., Daniels,D.L. and Blattner,F.R. (1993) *Genomics*, 16, 551–561.
3. Landy,A. (1989) *Annu. Rev. Biochem.*, 58, 913–949.

4. Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982) Molecular Cloning: A Laboratory Manual. Cold Spring Harbor University Press, Cold Spring Harbor.
5. Hsu,P.-L., Ross,W. and Landy,A. (1980) *Nature*, 285, 85–91.
6. Dunn,I.S. and Blattner,F.R. (1987) *Nucleic Acids Res.*, 15, 2677–2698
7. Daniels,D.L. (1990) In Drlica,K. and Riley,M. (eds.) The Bacterial Chromosome. American Society for Microbiology, Washington, D.C., pp. 43–52.
8. Bankier,A.T., Weston,K.M. and Barrell,B.G. (1987) *Methods in Enzymology*, 155, 51–93.
9. Hanahan,D. (1985) In Glover,D.M. (ed.) DNA cloning—A Practical Approach. IRL Press, Oxford. Vol.1, pp. 109–135.
10. D.L. Daniels and V. Burland, unpublished data.
11. Olson,C.H., Blattner,F.R.and Daniels,D.L. (1991) *Methods*, 3, 27–32.
12. United States Biochemical Corporation, 1990. Sequenase protocol booklet.
13. Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, 74, 5463–5467.
14. Daniels,D.L., Marr,L., Brumley,R.L. and Blattner,F.R. (1990) In Sarma,R.H., and Sarma,M.H. (eds.), Structure and Methods. Adenine Press, Guilderland, NY, Vol.1, pp. 29–35.
15. Lee,C.L., Gumport,R.I. and Gardner,J.F. (1990) *J. Bacteriol.* 172, 1529–1538.
16. Kieleczawa,J., Dunn,J.J. and Studier,F.W. (1992) *Science*, 258, 1787–1791.
17. Kotler,L.V., Zevin-Sonkin,D., Sobolev,I.A., Beskin,A.D. and Ulanovsky,L.E. (1993) *Proc. Natl. Acad. Sci. USA*, 90, 4241–4245.
18. Siemieniak,D.R., and Slightom,J.L. (1990) *Gene*, 96, 121–124.
19. Messing,J., Groneuborn,B., Muller-Hill,B. and Hofschneider,P.H. (1977) *Proc. Natl. Acad. Sci. USA*, 74, 3642–3646.
20. Yanisch-Perron,C., Vieira,J. and Messing,J. (1985) *Gene*, 33, 103–119.
21. Smith,V.,and Chee,M. (1991) *Nucleic Acids Res.*, 19, 6957.
22. Mizuuchi,M., and Mizuuchi,K. (1985) *Nucleic Acids Res.*, 13, 1193–1208.
23. Barnes,W.M., Bevan,M. and Son,P.H. (1983) *Methods in Enzymology* 101, 98–122.
24. Adachi,T., Mizuuchi,M., Robinson,E.A., Appella,E., O'Day,M., Gellert,M. and Mizuuchi,K. (1987) *Nucleic Acids Res.*, 15,771–784.
25. Liu,L., Whalen,W., Das,A. and Berg,C.M. (1987) *Nucleic Acids Res.*, 15, 9461–9469.
26. Ohara,O., Dorit,R.L.,and Gilbert,W. (1989) *Proc. Natl. Acad. Sci. USA*, 86, 6883–6887.
27. Innis, M.A., Myambo, K.B., Gelfand, D.H. and Brow, M.A.D. (1988) *Proc. Natl. Acad. Sci. USA*, 85: 9436–9440.
28. Shyamala,V., and Ames,G.F. (1989) *Gene*, 84, 1–8.
29. Strauss,E.C., Kobori,J.A., Siu,G. and Hood,L.E. (1986) *Analytical Biochem.*, 154, 353–360.